JP920000043

SPATIAL DATA MINING METHOD, SPATIAL DATA MINING APPARATUS AND STORAGE MEDIUM

Field of the Invention

The present invention relates to the processing performed
5 for the spatial data mining of databases, and more specifically, relates to a method and an apparatus for the calculation of optimal distances or optimal orientations, which is the basic function of spatial data mining.

Background of the Invention

10 A new processing technique has been introduced whereby spatial information, such as address data, in large databases can be interpreted by applying spatial context and spatial rules. But since present day spatial data mining requires expensive spatial/geometrical calculations
15 involving a huge amount of data, and since extremely difficult technical problems are frequently encountered, spatial data mining has not been well studied and remains an underdeveloped field. However, spatial data mining is considered to be a feasible basic technique that can greatly
20 assist in the development of databases for the information industry or for the GIS (Geographical Information System) field which have huge volumes of business. Spatial data mining, and associated techniques, is further considered to be a field having the potential to provide many benefits for
25 businesses.

1

Conventional spatial data mining systems, used for determining distances in advance through the introduction of correlated spatial rules, are well known. According to a method proposed by J. Han, et. al ("Spatial Data Mining:

5 Progress and Challenges", SIGMOD '96 Data Mining Workshop, pp. 55-69, 1996), for example, distance predicate terms "close to" and "far from" are defined, and correlated spatial rules, including the following two, are introduced from a spatial information database:

10 "close to a park" → "residential area" (support rate 5%, confidence rate 80%)

"drop in land price" → "far from a station" (support rate 10%, confidence rate 70%)

Further, another conventional spatial data mining system for

15 determining an orientation rule in advance for the introduction of a correlated spatial rule is also well known. According to the above method proposed by J. Han, et. al, spatial orientation terms "west of" and "north of" are defined, and correlated spatial rules, including the

20 following spatial orientation predicated ones, can be introduced from a spatial information database:

"west of a park" → "residential area" (support rate 5%, confidence rate 80%)

"drop in land price" → "north of a station" (support rate

25 10%, confidence rate 70%).

2

However, "close to" and "far from", which are included in the method proposed by J. Han, et. al, must be defined before data mining is initiated by providing a distance, such as "close to X" = "within a distance Y of X" and "far

5 from X" = "farther than a distance Z from X". In addition, "west of" and "north of" must be defined before data mining is initiated by providing a range and an angle, such as "west of X" = "the inside of a rectangle one side of which, to the west of X, has a length of Y" and "north of X" = "an

10 angle of $Y1°$ to $Y2°$ from X". At this time, a distance such as Y or Z, which is used for optimizing a specific objective function, or a numerical value for strictly determining an angle such as $Y1°$ or $Y2°$, which is used for optimizing a specific objective function, is requested by many

15 analyzation businesses, and even when the latest conventional techniques are employed, many of those analyzation businesses can not satisfactorily cope with their operation.

Conventional data mining systems can not, for example, cope

20 with a search for "a radius extending outward from a convenience store used to maximize the installation density of automatic teller machines within a unit distance in a district A" or a search to ascertain "the orientation of a route along which heavy air pollution spreads from a garbage

25 disposal area".

3

Summary of the Invention

To resolve the above technical shortcomings, it is one object of the present invention to provide a technique for calculating a distance or an orientation requested by 5 analyzation businesses. This technique is different from deriving a correlated spatial rule using distances or orientations which are calculated in advance.

It is another object of the present invention to increase the speed of the spatial data mining processing performed to 10 obtain a distance or an orientation.

It is an additional object of the present invention to provide spatial data mining output results that are useful to users (clients).

To achieve the above objects, according to the present 15 invention, a spatial data mining technique is provided that does not specify distances or spatial orientations in advance in order to introduce a spatial correlative rule, but instead, employs as input parameters the definition of a distance or an orientation, the definition of a set of 20 starting points and the definition of an objective function for obtaining a distance or an orientation that is requested by many analyzation businesses, and that is used for optimizing a specific objective function. Specifically, according to the present invention, a spatial data mining 25 method, for introducing spatial rules from a database in

which spatial information, such as addresses, is stored, comprises the steps of: providing from the database a starting point or a starting point group; employing the starting point or the starting point group to define a

5 distance or an orientation; defining an objective function that is examined in order to introduce a spatial rule; and calculating a distance from or an orientation block originating at the starting point or the starting point group in order to optimize the objective function that is

10 defined.

The objective function is a function for which a distance or an orientation requested by an analyzation business is not provided. The spatial data mining method further comprises a step of: entering as input parameters the definition of a

15 distance, the definition of the starting point or the starting point group and the definition of the objective function.

At the step of calculating the distances, an intermediate table is generated based on starting point set data

20 consisting of the starting point group and the objective function, and in accordance with distance values, attribute values for query points in the database are added together, based on the intermediate table. As a result, the calculation time can be considerably reduced.

25 The spatial data mining method further comprises a step of: displaying on a map the distance or the orientation block

5

relative to the starting point or the starting point group. Then, a user can visually identify the rule acquired by performing the calculation, so that for the user the usability is improved.

5 The orientation block can be obtained by employing the numerical value of the orientation used to optimize the objective function. Further, a search objective data range, at equal distances from the starting point and from the starting point group, that is appropriate for calculating an 10 orientation can be selected as the orientation block. The calculation from the starting point or the starting point group for an infinite range is almost impossible, and determination of an optimized area is effective.

According to the present invention, a spatial data mining 15 method, for generating a data table used to introduce a spatial rule for the orientation obtained from a spatial information database, comprises the steps of: providing a set of starting points and a set of query points in a database; designating an upper limit for a distance between 20 the set of starting points and the set of query points; calculating a distance between each starting point and each query point; calculating an angle formed between a starting point and a query point whose distance from the starting point does not exceed the designated upper limit; and 25 generating a data table using the angle formed with the starting point. The query points can be a set of points representing customer data, and are employed to actually

6

calculate a distance from or an orientation relative to a starting point or starting point groups.

Furthermore, according to the present invention, a spatial data mining apparatus for calculating an optimal distance
5 from a database, wherein spatial information, such as addresses, is stored, comprises: input means, for the input of an objective function required for the optimization of a distance; intermediate table generation means, for employing in the database starting point data and query point data for
10 calculating the distances between each starting point and each query point and for generating an intermediate table; and optimal distance calculation means, for calculating a distance, based on the intermediate table generated by the intermediate table generation means, in order to optimize
15 the value of the objective function that is entered by the input means.

The intermediate table generation means includes: Voronoi diagram preparation means, for preparing a Voronoi diagram by using the starting point data in the database; distance
20 calculation means, for employing the Voronoi diagram, prepared by the Voronoi diagram preparation means, and the query point data in the database to calculate distances between individual starting points and individual query points and to generate data records; and individual distance
25 calculation means, for selecting an optimization function from among objective functions to be examined, and for adding together record values, collected from the data

records, that are required for optimization of each of the distances.

Further, the Voronoi diagram preparation means repeats plane quarter division in accordance with the number of starting
5 points that are entered, sorts the starting points into end plane pixels obtained by division and selects one starting point in each of the end plane pixels as a representative point for the pertinent pixel, prepares a quaternary incremental tree with pixels at individual levels being
10 defined as intermediate nodes, scans the individual nodes of the quaternary incremental tree in the breadth-first order, beginning at the topmost level, and outputs a set of starting points that are positioned in ranks. As a result, high speed processing can be performed.

15 When the structure of the quaternary incremental tree is calculated in advance and stored in memory, the high-speed mining process for distance optimization or for orientation can be implemented because such a tree structure is frequently used for the mining process.

20 According to the present invention, a spatial data mining apparatus for calculating an optimal orientation for a database, which includes spatial information, such as addresses, comprises: input means, for the input of an objective function required for the optimization of an
25 orientation; intermediate table generation means, for employing, based on starting point data and query point data

in the database, angles of 0 degrees from the starting points in a specific direction to generate an intermediate table in which the orientation of the locations of the query points are included; and optimal orientation calculation
5 means, for calculating, based on the intermediate table generated by the intermediate table generation means, an orientation for optimizing the value of the objective function that is entered by the input means.

The intermediate table generation means includes: Voronoi
10 diagram preparation means, for preparing a Voronoi diagram by using the starting point data in the database; distance calculation means, for employing the Voronoi diagram prepared by the Voronoi diagram preparation means and the query point data in the database to calculate distances
15 between individual starting points and individual query points; orientation calculation means, for calculating, based on the distances obtained by the distance calculation means, orientations of the starting points with the query points that fall within a designated distance upper limit,
20 and for storing the orientations as data records for the intermediate table; and individual orientation calculation means, for selecting an optimization function from among objective functions to be examined, and for collecting and adding record values, from the data records, that are
25 required for optimization of each of the orientations.

According to the present invention, a spatial data mining apparatus, for calculating an optimal distance from or an

optimal orientation with a database in which spatial information, such as addresses, is stored, and for outputting the optimal distance or the optimal orientation, comprises: input means, for the input of an objective
5 function for which a distance or an orientation requested by an analyzation business is not provided; optimal distance/orientation calculation means, for employing starting point data and query point data in the database for calculating a distance between, or the orientation of each
10 of the starting points with each of the query points, and for calculating the optimal distance or the optimal orientation for the optimization of the value of the objective function; and display means, for displaying, on the screen of a geographical information system, the optimal
15 distance or the optimal orientation calculated by the optimal distance/orientation calculation means.

The display means can use the optimal distance calculated by the optimal distance/orientation calculation means for the display of circular areas, the centers of which are starting
20 points. The display means can also use the optimal orientation, calculated by the optimal distance/orientation calculation means, for the display of fan-shaped portions of the circular areas, the origins of the fan-shaped portions being the starting points at the centers of the circular
25 areas. Thus, for easy understanding, the obtained optimal distance/orientation can be displayed on maps, so that customer usability can be considerably improved.

According to the present invention, a spatial data mining apparatus, for introducing a spatial rule from a database, which also includes spatial information, such as addresses, comprises: starting point provision means, for providing

5 starting points or starting point groups obtained from the database; objective function definition means for defining an objective function that is to be examined in order to introduce a spatial rule; distance calculation means, for calculating distances originating at the starting points or

10 at the starting point groups for optimizing the objective function that is defined; orientation definition means, for employing the starting points or the starting point groups to define distances or orientations; and orientation block calculation means for calculating orientation blocks

15 beginning at the starting points or the starting point groups to optimize the objective function that is defined.

The spatial data mining apparatus further comprises: starting point/query point provision means for providing a set of starting points and a set of query points in the

20 database; distance upper limit designation means for designating the upper limit for a distance between the set of starting points and the set of query points; distance calculation means for calculating a distance between each starting point and each query point; angle calculation means

25 for calculating an angle formed between a starting point and a query point whose distance from the starting point does not exceed the designated upper limit; and a data table generation means for generating a data table using the angle

11

formed with the starting point.

According to the present invention, a storage medium is provided on which is stored a spatial data mining program, which introduces a spatial rule extracted from a database

5 that includes spatial information, such as addresses, based on an objective function for which neither a distance nor an orientation is provided, the program comprising the steps of: providing a starting point or a starting point group from the database; employing the starting point or the

10 starting point group to define a distance or an orientation; defining an objective function that is to be examined; and calculating a distance measured from the start point or the starting point group, or an orientation block to optimize the objective function that is defined. The storage medium

15 can be a portable medium, such as a CD-ROM, or can be a storage medium such as a hard disk at a program provider on which programs are stored for downloading via a network or a hard disk a user employs to store programs that are so downloaded.

20 Brief Description of the Drawings

Fig. 1 is a diagram showing a first modeling example output by a distance optimization engine according to one embodiment of the present invention.

Fig. 2 is a diagram showing a second modeling example output

25 by an orientation optimization engine according to the

12

embodiment of the present invention.

Fig. 3 is a flowchart for explaining an overview of the algorithm of the distance/orientation optimization engine according to the embodiment.

5 Fig. 4 is a diagram showing an example database.

Fig. 5 is a diagram showing example definitions for a staring point group and a reference group used as input parameters.

Fig. 6 is a diagram showing example definitions for a 10 distance and an orientation.

Fig. 7 is a diagram showing an example definition for an objective function.

Fig. 8 is a diagram showing processing for which an incremental method is used.

15 Fig. 9 is a diagram showing the processing for a Voronoi diagram for which a point position determination method is used.

Fig. 10 is a diagram showing the pre-process for the incremental method.

20 Fig. 11 is a schematic diagram for explaining the

configuration of a computer system used as a spatial data mining apparatus.

Fig. 12 is a block diagram for explaining the configuration of the spatial data mining system according to the embodiment.

Fig. 13 is a block diagram for explaining the arrangement of a starting point adding order determiner in a Voronoi diagram generator of Fig. 12.

Fig. 14 is a diagram for explaining divided plane pixels and a quaternary incremental tree that is generated.

Fig. 15 is a block diagram for explaining the structure of a Voronoi diagram addition unit in the Voronoi diagram generator in Fig. 12.

Detailed Description of Invention

A preferred embodiment of the present invention will now be described in detail while referring to the accompanying drawings.

First, modeling and an algorithm used for the present invention will be described so that the spatial data mining in this embodiment can be easily understood.

Fig. 1 is a diagram showing a first modeling example output

for this embodiment by a distance optimization engine. On a map 11, distances (or distance blocks), for the optimization of an objective function, measured from starting points (or a starting point group) 12 to 14, which are convenience

5 stores (CS), constitute the predetermined radii of circles 15 to 17. In the example in Fig. 1, the displayed distances, measured from the starting points 12 to 14, used for the optimization of an objective function were obtained by a search run to determine the "distances from convenience

10 stores within which the bag-snatching occurrence rate is maximized."

In this example for the embodiment, the content output is:

"bag-snatching" → ("convenience stores", "[0,100]"), "five" cases.

15 This means that at each "convenience store", within a radius of between "0"m and "100"m, the "bag-snatching" occurrence rate is maximized, and that "five" cases occurred within the radius (m).

The examples output for the distances from the starting 20 points are as follows:

"bag-snatching" → ("station", "[50, 180]"), "6.1" cases
"bag-snatching" → ("banks", "[200, -]"), "2.2" cases
"burglary" → ("banks", "[60, 200]"), "1.3" cases
"murder" → ("restaurants", "[0, 50]"), "0.4" cases

15

Fig. 2 is a diagram showing a second modeling example output for this embodiment by an orientation optimization engine. On the map 11, orientation blocks 18 to 20, for the optimization of the objective function, relative to the

5 starting points (or the starting point group) 12 to 14, the convenience stores (CS), are displayed on the map 11. In the example in Fig. 2, the angles for the fan-shaped areas that are shown were provided by a search conducted to determine "the orientations for areas at the convenience

10 stores within which the bag-snatching occurrence rate is maximized".

In this example, the content output is:

"bag-snatching" → ("convenience stores", "[0, 100]"), "five" cases.

15 This means that within an area having an orientation of "0"° to "100"° the "bag-snatching" occurrence rate is maximized and that for every 10° "five" incidents of "bag-snatching" occurred.

Another example output for the orientation of an area

20 originating at a starting point is:

"bag-snatching" → ("shrines", [120, 240]"), "6.1" cases.

Fig. 3 is a flowchart for explaining an overview of the

algorithm used for the distance/orientation optimization engine in accordance with the embodiment. First, a starting point or a starting point group is provided (step 101). That is, for the optimization of a distance, an entity 5 (substance) on a map is designated as a starting point for the calculation of the distance, while to obtain a specific starting point for orientation optimization, an entity on a map is designated an orientation reference point. Either a single or multiple points, defined as a starting point or a 10 set of starting points, may be employed for each calculation. Then, the distance from the starting point (or the starting point group) or the orientation, for which the location of the starting point is provided, is defined (step 102). It should be noted that if the Euclidean distance is 15 defined as distance, no distance definition is required. An objective function is defined for a desired objective, such as the total sales of a specific product or the number of crimes that occurred (step 103). Thereafter, for the optimization of the objective function, calculations are 20 performed for the arrangement, relative to the starting point, of a distance block or an orientation block (step 104). Following this, distance/orientation buckets are collected and added together, and the optimal distance is calculated. When n denotes the number of queries while m 25 denotes the number of starting points, and N = n + m, the average time required for the distance calculation can be represented as O(N). That is, in order to find a starting point corresponding to a query point, the average number of calculations is O(logn), distance is obtained by calculating

O(1), and the time required for this process is O(nlogn).
When the Euclidean distance is defined, however, the average
processing time is determined to be O(n) by using a
quaternary incremental calculation method that will be
5 described later. The query point is customer data stored in
a target database used for processing.

Fig. 4 is a diagram showing an example database. In this
embodiment, it is assumed an integrated geographical
information system, which is associated with a database
10 having the schema as shown in Fig. 4, is present. Each
schema for the database includes ID information for
identifying data and position (coordinate) information. The
position information includes address data and coordinate
information corresponding to map information. In addition,
15 the underlined numerical attributes and the underlined
categorical attributes printed in italics are also included.
The mining using the optimization rule for the distance and
the orientation is performed in the database.

Fig. 5 is a diagram showing as input parameters example
20 definitions for a starting point group and a reference point
group (starting point group for the orientation). In this
example, the starting point group for the distance and the
starting point group for the orientation at step 101 in Fig.
3 are defined. Also in this example, entities, such as post
25 offices, schools and police stations, are defined and
designated as starting points and reference points. For
example, post offices (ALL) are defined by totaling a

plurality of categorical attributes, such as mail collection and delivery offices and privately-owned post offices. Similarly, schools (ALL) are defined by totaling the categorical attributes, such as various special schools,

5 grade schools and junior high schools. For stations (ALL), a starting point group or a reference point group is defined based on numerical attributes, such as equal to or greater than X number of customers or less than X number of customers. Likewise, convenience stores can be defined

10 based on numerical attributes, such as sales equal to or greater than X or less than X.

Fig. 6 is a diagram showing definition examples for a distance and an orientation, which are performed at step 102 in the flowchart in Fig. 3. These definitions are

15 designated as input parameters for mining performed using the optimization rule. The distance definition is a Euclidean distance or a network distance. The Euclidean distance is calculated using a Voronoi diagram; however, for high-speed calculations or for a very short distance, the

20 obtained distance value would be dissociated from the object represented on a map. The network distance is obtained using the Dijkstra algorithm. In accordance with the Dijkstra algorithm, while the shortest distance to each node is obtained beginning at the periphery of the start node,

25 the range is gradually expanded, until finally, the shortest distance for all the nodes is obtained. An extended period of time is required for the calculations; however, the object represented on a map can be reflected. In this

embodiment, the calculation method used for the Euclidean distance is employed, which differs from the calculation method used for the other types of distances. No distance definition is required to obtain the Euclidean distance.

5 And for orientation, an orientation scale having one clockwise cycle of 360°, with north defined as 0°, is employed for defining an orientation block.

Fig. 7 is a diagram showing a definition example for the objective function, which is the definition given at step

10 103 in the flowchart in Fig. 3. The objective function is designated for mining for which the optimization rule is used. The objective function can be defined based on the underlined numerical (or derived as numerical values) attributes for the individual schema and the underlined

15 categorical (or derived as category values) attributes printed in italics. For example, as a customer schema, a numerical attribute can be used to define a "maximized distance for the 'average year income' of customers having a support rate of S or higher", or the categorical attribute

20 can be used to define a "maximized distance for the ratio of customers that are 'sixty years old or older' having a support rate of S or higher". In addition, as the ATM schema, the categorical attribute can be employed to define a "maximized distance for an 'ATM count/customer count'

25 obtained for customers having a support rate of S or higher."

At step 104 in the flowchart in Fig. 3, the distance block

20

from or the orientation block relative to the starting point for the optimization of the defined objective function is obtained. In this embodiment, an intermediate table is prepared for high-speed processing, as will be described later. For the intermediate table, a Voronoi diagram (Thiessen division), which is a geometric figure for which a set of starting points are provided as generating points, is prepared using the incremental method. That is, a Voronoi diagram V_{m+1} including m + 1 points is prepared by adding a new generating point P_{m+1} to a Voronoi diagram V_m in which generating points P_1 to P_m are included.

Fig. 8 is a diagram showing the processing performed by the incremental method. The processing is shown in the flowchart on the left, and the preparation procedure used for the Voronoi diagram is shown on the right. First, from among the points P_1 to P_m, the closest point P to P_{m+1} is obtained using the "fast point position determination method for the Voronoi diagram", which will be described later (step 111). Then, a vertical bisector L of PP_{m+1} is drawn in the Voronoi area of the point P (step 112). Thereafter, a vertical bisector is also drawn for the generating point in the Voronoi area contacted by the line L (step 113). This processing is repeated, and the Voronoi area for the point P_{m+1} is prepared and defined as V_{m+1} (step 114).

Fig. 9 is a flowchart showing the processing performed by the point position determination method for the Voronoi diagram. According to this method, in the Voronoi diagram

21

V_m, which includes the generating points P_1 to P_m, the closest generating point to a specific point P is obtained (a Voronoi area wherein the specific point P is located is obtained). First, a specific point Pi is selected from

5 among the points P_1 to P_m, and the distance d between the point P is obtained (step 121). Then the distance dj between the generating point Pj adjacent to the point Pi and the point P is calculated and compared with the distance d (step 122). When dj < d is determined (step 123), d = dj

10 and Pi = Pj are defined, and program control returns to step 122. When no such generating point is present, Pi is regarded as the generating point (step 125) and the processing is thereafter terminated.

Fig. 10 is a diagram showing the pre-processing performed

15 using the incremental method. First, a quaternary incremental tree having a depth d is prepared wherein a generating point is substantially included in each pixel (step 131). Then, numbers are provided for the pixels as is shown in Fig. 10(a) (step 132). Shown in Fig. 10(a) is a

20 view of a map (two-dimensional plane) on which arrows are used to indicate the order used for providing pixel numbers. The generating points are allocated for the pixels (the leaves of the quaternary incremental tree) in accordance with coordinate values, and labels are provided for the

25 leaves (step 133). The labels of the leaves are copied for all the ancestors (non-labeled) of the leaves (step 134). Finally, the sections and leaves of the quaternary incremental tree are arranged in the breadth-first order,

which is used for the "incremental method" that is employed (step 135). A view of a quaternary incremental tree having a depth of 3 is shown in Fig. 10(b), and for this tree the order used for the incremental method is not that which

5 corresponds to the depth, but is that which corresponds to the transverse direction. It should be noted that the Voronoi diagram may be prepared in advance for the entities on a map, such as train stations, post offices, police stations, schools and parks, that tend to be analysis

10 targets and that do not move.

The distance from the starting point using the Voronoi diagram is calculated. For this processing, the "resolving of a point position determination problem for n query points using a Voronoi diagram V_m consisting of m generating points

15 P_1 to P_m" is performed. That is, the distance between each of n (a considerable number) query points (e.g., crime data) and each of a set of m starting points (convenience stores) is calculated. It should be noted that the obtained distance is the distance from the closest starting point.

20 More specifically, the first query point at each pixel is calculated by using a generating point that serves as a label for the pertinent pixel at the succeeding point. The other query point at each pixel is calculated by using the generating point that is determined to be the closest point

25 using the preceding distance/orientation calculation. In this manner, the intermediate table is generated.

Finally, the buckets for the distances or orientations are

23

collected and added, and the optimal distance is calculated. For this processing, the attribute values of the individual n (a considerable number) query points are collected and added in accordance with the distance/orientation value.

5 That is, the "data count" and the data required for the calculation of the objective function value are collected for each bucket. For example, for the objective function of a "customer schema" (the maximized distance for the "average yearly incomes" of customers having a support rate of S or

10 higher), a data count and a total annual income value that are consonant with the distance are sequentially added, and the aggregate is output. When the aggregate information is swept once, the optimized distance is obtained. This result is displayed as a circle or as a fan-shapes area on the map,

15 as is shown in Fig. 1 or 2.

This completes the explanation for the algorithm for the processing performed in this invention. The processing algorithm can be provided and executed as a computer program.

20 Fig. 11 is a schematic diagram for explaining the arrangement of a computer system that constitutes a spatial data mining apparatus. The processing algorithm of the embodiment can also be provided as a program to be executed by the computer system in Fig. 11. The processing program

25 is stored on a hard disk drive (HDD) 75, is loaded into a main memory 72 for execution, and is executed by a CPU 71. The HDD 75 also holds a large database, including such

24

spatial information as addresses, and the processing program is used to access the database. Geographical information obtained by the geographical information system (GIS) and the distance or the orientation optimized using calculations

5 are provided for a user by a display device 76. The user employs an input device 77 to enter desired objective functions or data output commands. The input device 77 includes a keyboard, a mouse, a pointing device or a digitizer. The output results can be stored on a floppy

10 disk loaded in a floppy disk drive (FDD) 73, an auxiliary storage device, from which new data can also be obtained. Further, a CD-ROM drive 74 can also be employed for data input.

The computer program that implements the processing

15 algorithm of this embodiment can be stored on a storage medium, such as a floppy disk or a CD-ROM, which may be carried by a user. In this case, the data extraction section of an ordinary database search-program or a program provided only for the display of data on the display device

20 76 may be stored on the HDD 75 in advance. Therefore, it is normal for other sections to be distributed using various types of storage media. A communication device (not shown) may be connected to a bus 78, so that a remote database can be employed to perform the processing, or so that the

25 processing results can be transmitted to a remote area. That is, a large database in which spatial information, such as addresses, is included can also be provided outside the configuration shown in Fig. 11.

The configuration of the embodiment will now be described in detail while referring to a functional block diagram.

Fig. 12 is a block diagram for explaining the configuration of a spatial data mining system according to this
5 embodiment. In Fig. 12, the arrangement of the CPU 71 in Fig. 11 is shown in detail. The CPU 71 mainly comprises an intermediate table generator 30 and an optimal distance calculator 39. The intermediate table generator 30 includes a Voronoi diagram generator 31, a distance calculator 32 and
10 a record calculator 33. The Voronoi diagram generator 31 receives data for a set of starting points, such as a set of points for convenience stores, that consists of IDs, names and coordinates on a map, and employs a start point adding order determiner 34 and a Voronoi diagram point addition
15 unit 35 to generate a Voronoi diagram. The distance calculator 32 receives data for a set of query points, such as a set of points representing customer data, that consists of IDs, names, coordinates on a map and payment values, and generates a customer data record, or a set of customer data
20 records, for which distance is obtained. The record calculator 33 employs the customer data record output by the distance calculator 32 to collect and add record values required for the optimization of each distance. As a result, an intermediate table shown in Fig. 12 is prepared.
25 In this table, the record value and the total payment value are shown for each distance block. The optimal distance calculator 39 scans the intermediate table to obtain the

distance at which the best objective function value is obtained, and outputs this distance as the distance optimization rule.

The functions of the starting point adding order determiner
5 34 of the Voronoi diagram generator 31 will now be described while referring to Figs. 13 and 14.

Fig. 13 is a block diagram for explaining the configuration of the starting point adding order determiner 34 of the Voronoi diagram generator 31. The starting point adding
10 order determiner 34 includes a plane quarter division function 41, a starting point distribution function 42, a representative point determination function 43, a quaternary incremental tree generation function 44 and an adding order determination function 45. The plane quarter division
15 function 41 receives all the data for a set of starting points, and repeats the plane quarter division t ($t = m^{1/2} - 1$) times, in accordance with the input starting point count m. That is, while one pixel at each level is divided by four, it is substantially preferable that each pixel include
20 one starting point.

Fig. 14 is a diagram for explaining divided plane pixels and a quaternary incremental tree that is prepared. As is shown in Fig. 14, the numbers are provided for pixels divided by the plane quarter division function 41. In this example, 64
25 pixels 0 to 63 are obtained using a depth of 3. The starting points distribution function 42 in Fig. 13

allocates m input starting points to the smallest divided
planes (0 to 63). In Fig. 14, starting points 8 (S8) and 19
(S19) are allocated for a pixel 0, and a starting point 12
(S12) is allocated for a pixel 60. The representative point
5 determination function 43 in Fig. 13 selects one of these
starting points in each pixel as the representative point
for the pertinent pixel. For example, the starting point 12
(S12) is selected for the pixel 60 in Fig. 14. When
multiple starting points are present, an arbitrary point is
10 selected (e.g., a starting point 8 (S8) is selected for the
pixel 0). When there is no starting point, no
representative point is selected.

The quaternary incremental tree generation function 44 in
Fig. 13 prepares a quaternary incremental tree wherein the
15 pixels at individual levels (not the lowermost level), which
appear during the repetitive quarter division process, serve
as intermediate nodes (see the diagram on the right in Fig.
14). Furthermore, the representative points of the
intermediate nodes are determined beginning with the
20 intermediate node at a lower level of the tree. In Fig. 14,
one of the representative points of the child nodes for the
individual intermediate nodes is regarded as the
representative point, and the starting point 8 (S8) is
sequentially selected. The adding order determination
25 function 45 in Fig. 13 scans the nodes of the quaternary
incremental tree in the breadth-first order, beginning with
the node at the topmost level. In other words, the nodes
are not scanned along the depth (the depth priority order),

28

but transversely at each level. During this process, a representative point that is not yet on the output starting point list for each node is added at the last of the list. For the leaf node at the lowermost level, not only the

5 representative point, but also all the starting points that belong to the pertinent node and that have not yet been entered on the list are added. That is, in the example in Fig. 14, the starting point 19 (S19) for the pixel 0 is added. Through this processing, the set of numbered

10 starting points is output by the starting point adding order determiner 34 to the Voronoi diagram point addition unit 35 in Fig. 12.

Fig. 15 is a block diagram for explaining the arrangement of the Voronoi diagram point addition unit 35 of the Voronoi

15 diagram generator 31 in Fig. 12. The Voronoi diagram point addition unit 35 mainly includes a fast point position determination function 51 and an area division function 52. The fast point position determination function 51 has a starting position determination function 53 and an

20 asymptotic function 54, and obtains, among points P_1 to P_m, the closest point P to the point P_{m+1} provided by the "fast point position determination method". The starting position determination function 53 obtains a pixel X, at the lowermost level of the quaternary tree, to which an input

25 point belongs, and defines, as the starting point, the point (Voronoi point) P_i that is the representative point of the pixel X. If no representative point is selected, the representative point of the parent node is defined as the

starting point. The asymptotic function 54 calculates the
distance d between the point Pi and the input point P_{m+1}.
The distance dj between the point P_{m+1} and each Voronoi point
Pj adjacent to the point Pi is then calculated, and is

5 compared with the distance d. If dj < d, d = dj and Pi = Pj
are determined. If such an adjacent generating point is not
present, Pi is determined to be the starting point, and the
representative point for the pixel X is defined as the point
Pi. As is shown on the right in Fig. 15, the area division

10 function 52 draws a vertical bisector L of PP_{m+1} in the
Voronoi area of the point P. Further, when the starting
point of the Voronoi area that the bisector L contacts is
defined as a point P, a vertical bisector is also drawn.
This process is repeated until an area is defined by

15 vertical bisectors, so that the intermediate Voronoi diagram
V_{m+1}, which includes m + 1 points, is obtained as is shown on
the right in Fig. 15.

The intermediate table generator 30 in Fig. 12 outputs
customer data records for which the distance is obtained, in

20 the above described manner, by the distance calculator 32,
which uses the data for a set of query points and the
Voronoi diagram that is thus generated by the Voronoi
diagram generator 31. That is, using the fast point
position determination method, the Voronoi point closest to

25 the input point is extracted from the query point record for
the query point set and the intermediate Voronoi diagram
V_{m+1}. Then, the distance between the query point record and
the Voronoi point is obtained, and the query point record is

output with the obtained distance.   Thereafter, as was described above, the intermediate table, for which the record values are collected and added for each distance block by the record calculator 33, is output by the 5 intermediate table generator 30.   The information for the intermediate table is stored on HDD 75, for example.

The optimal distance calculator 39 employs this intermediate table to calculate the optimal distance.   For example, for the customer data in the query point set data in Fig. 12, 10 the payment value is accumulated in order, beginning with the record for the smallest distance, and during this process, the distance that provides the highest "accumulated payment values/accumulated records" value is recorded.   At this time, generally, during the accumulation process, an 15 intermediate value located between the record providing the maximum value and the distance value for the next record in the intermediate table is recorded as a temporary optimal distance.   As is described above, while the temporary optimal distance that provides the maximum value for the 20 objective function is maintained, the intermediate table is scanned once, so that the optimal distance can be obtained. The obtained optimal distance can be displayed by the display device 76, as is shown in Fig. 1.

As is shown in Fig. 2, only the data inside each circle that 25 are required for the optimization are included in the intermediate table of the orientation optimization algorithm.   That is, the distance from the reference point

31

is calculated for each of the records (query points) in the database that are associated with the spatial information, and the angle formed by a reference point and a query point that falls within the radius of the circle is calculated.

5 The distance is entered in the record for the intermediate table in accordance with the obtained angle value. If circles required for the search overlap each other, a query point may be included in multiple circles at the same time. Thus, to prepare the intermediate table, one of the

10 following two methods is employed:

* An angle based on each reference point is added to all the corresponding records in the intermediate table.
* Only an angle relative to the reference point located at the shortest distance is added to a corresponding record in

15 the intermediate table.

The orientation optimization intermediate table includes as attributes, at equal intervals, angle intervals in the ascending or descending order. For example, when the average annual income is employed as the objective function,

20 the intermediate table includes the total annual income and the record count required for this calculation.

total annual income	angle interval	record count
8,400,000	0 to 10	6
5,000,000	10 to 20	1
7,000,000	20 to 30	3
5,800,000	30 to 40	3

32

Thereafter, the optimal orientation range for the objective function is determined using the intermediate table. For example, when the indexes of the records in the ascending order of the angles are defined as s and t, the average 5 income X[s, t] for the interval [s, t] is represented as

X[s, t] = "the total of incomes of the records included in the interval [s, t] ÷ the number of salary income persons in the interval [s, t]".

This, then, is a problem for obtaining the interval [s, t] 10 for optimizing X[s, t]. While the interval algorithm can be performed using O(n), where n denotes the intermediate table size, in this embodiment, an algorithm is employed that takes into account the discontinuity at the degree of 0 that occurs when an angle is represented as a numerical value. 15 For example, for records that are sorted in the intermediate table, the annual income and the number of records are accumulated, and:

* a record position at which a Σ annual income/a Σ record count is maximized, and the Σ annual income and the Σ record 20 count at the position t, and

* a record position at which a Σ annual income/a Σ record count is minimized, and the Σ annual income and the Σ record count at the position s

are stored. The optimal orientation can be acquired by

33

using the value obtained by performing a scan of the table of up to 360 degrees. When s < t, the optimal interval is provided as [s, t], with no 0 in between, and the average annual income can be obtained quickly. If t < s, the

5 optimal interval is [s, t], with a 0 in between, and the average annual income can be obtained from the total number of query points and the total of the annual incomes relative to the query points. The obtained results can then be displayed on a map by the display device 76, as is shown in

10 Fig. 2.

According to the embodiment of the present invention, the definition of the distance (either the Euclidean distance (substantially, the linear time in this case) or the network distance (the polynomial time in this case)), the definition

15 of the orientation (direction), the definition of a set of starting points, and the definition of the objective function are designated as input parameters, as needed. Thus, the optimization rules can be calculated and listed. For example, if the following is designated:

20 distance: Euclidean distance
definition of a set of starting points:
post offices (ALL, each type), schools (ALL, each type)
definition of an objective function:
customers (maximized distance for average annual incomes of

25 customers having a support rate of S or higher)
customers (maximized distance for amount of mutual information for each sex),

34

the optimization rules for the objective functions from the set of starting points are, for example:

"within X from post offices to maximize the average annual income (support rate s, average annual income x)" and

5 "within X from universities to maximize mutual information amount for each sex (support rate s, entropy gain g)"

are listed in accordance with the combination of the above definitions. Thus, in accordance with the data in which a user is interested, the user can sort or file these rules.
10 In addition, for a matter of special interest, the optimization rule can be displayed on the map by the GUI, as is shown in Fig. 1 or 2.

As is described above, according to the present invention, a spatial data mining technique can be provided for obtaining
15 a distance or an orientation requested by many analyzation businesses, without having to define the distance or the orientation in advance in order to introduce a spatial correlative rule.